CLAIMS

1 A method for producing a nucleic acid library,

which library contains a plurality of different nucleic acid fragments, the combination of said fragments being a representative partition of the entirety of a sample nucleic acid,

the method comprising:

(i) digesting the sample nucleic acid with a plurality of different restriction enzymes to generate a plurality of different layers of fragments,

wherein each layer is a group of fragments having a unique combination of restriction ends,

and wherein the combination of layers represents the entirety of the sample nucleic acid,

(ii) optionally purifying said fragments,

(iii) selecting a desired sub-set of layers according to the unique restriction ends of said layers,

(iv) ligating said sub-set of layers into vectors adapted to receive it,

(v) transforming host cells with the vectors

(vi) culturing said host cells to provide said library containing said partition of the sample nucleic acid.

2 A method as claimed in claim 1 wherein the sample is genomic DNA.

3 A method as claimed in claim 2 wherein the sample consists of an entire genome.

4 A method as claimed in any one of the preceding claims wherein the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) are selected in order to generate a library size with a reduced complexity compared to the sample nucleic acid of at least 10, 100, or 1000-fold.

5 A method as claimed in any one of the preceding claims wherein between 3 and 6 restriction enzymes are used.

6 A method as claimed in any one of the preceding claims wherein the digestion by one restriction enzyme is partial, and the group of fragments in the selected layer have restriction ends created by said partial digestion.

7 A method as claimed in any one of the preceding claims wherein the selected sub-set of layers consists of one layer.

8 A method as claimed in any one of claims 1 to 6 wherein the sub-set of layers consists of two layers.

9 A method as claimed in any one of the preceding claims wherein the fragments are purified at step (ii).

10 A method as claimed in claim 9 wherein the purification removes fragments of less than 100 bases.

11 A method as claimed in any one of the preceding claims wherein the size range of the fragments in the library is between 100 and 2000 bps.

12 A method as claimed in any one of the preceding claims wherein enhancement linkers are added prior or during step (iv) to prevent undesired sub-sets of layers being included in said library,

each of which enhancement linkers comprises:

(i) a core sequence,

(ii) a portion that matches the restricted-end of an undesired sub-set, and

(iii) a sequence to inhibit the fragments in the undesired sub-set recombining.

13 A method as claimed in claim 12 wherein the enhancement linkers comprise any of those given in Table 1.

14 A method as claimed in any one of the preceding claims wherein adaptor oligonucleotides are used in step (iv) to facilitate the ligation of the desired sub-set of layers into vectors adapted to receive it.

15 A method as claimed in any one of the preceding claims wherein said sample is derived from one of the following organisms or species : Human, Arabidopsis, wheat, rice, millet, soybean.

16 A method as claimed in any one of the preceding claims wherein libraries are prepared separately using methylation sensitive and non-sensitive restriction enzymes, whereby comparison of the libraries permits methylation distribution patterns in the sample to be revealed.

17 A method as claimed in any one of the preceding claims wherein the sequence of the sample nucleic acid is known, and the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) are selected to produce the desired library size in accordance with the restriction site frequency of each enzyme in the sample nucleic acid sequence.

18 A method as claimed in claim 17 wherein the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii), are selected in accordance with the formula:

$$N_{x1\sim x2} = GP_1^2 \sum_{k=x1}^{k=x2} \prod_{i=1}^{i} (1-P_i)^k$$

Nx1~x2 is the number of fragments with length between x1 and x2
k is fragment length
x1 and x2 are upper and lower limits of the size range of the fragments in the library
Pi is the probability of having a restriction site at any given base for the 'i'th enzyme.

19 A method as claimed in claim 17 or 18 wherein a representative partition of a particular region is produced in accordance with a restriction map of the sample nucleic acid sequence.

20 A method as claimed in any one of claims 1 to 16 wherein the size of the sample nucleic acid is known, and the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) are selected to produce the desired library size in accordance with an assumed restriction site frequency of each enzyme in the sample nucleic acid.

21 A method as claimed in claim 20 wherein the restriction site frequency within the sample is assumed based on sequence information from the sample.

22 A method as claimed in claim 20 wherein the restriction site frequency is assumed to be randomly distributed

23 A method as claimed in any one of claims 20 to 22 wherein the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii), are selected in accordance with the formula:

$$N_{x1\sim x2} = GP_1^2 \sum_{k=x1}^{k=x2} \prod_{i=1}^{i} (1-P_i)^k$$

Nx1~x2 is the number of fragments with length between x1 and x2
k is fragment length
G is the size of the sample
x1 and x2 are upper and lower limits of the size range of the fragments in the library
Pi is the probability of having a restriction site at any given base for the 'i'th enzyme.

24 A method as claimed in claim 23 wherein the restriction enzymes used in step (i) are 4 and 6nt cutting restriction enzymes, and are selected on the basis of the formula:

$$N' = 4^{-12} v' G \int_{k=x1}^{k=x2} \sum \left[(1-1/4^4)^{nk} (1-1/4^6)^{(1+m)k} \right]$$

wherein:

k is fragment length
G is the size of the sample
x1 and x2 are upper and lower limits of the size range of the fragments in the library
n is the number of extra 4 nt cutters
m is the number of extra 6 nt cutters

25 A method as claimed in any one of claims 20 to 24 wherein the size of the resulting library is estimated by the further steps of:
(vii) sequencing the fragments in a fraction of the host cells in said library,
(viii) estimating the size of the library using formula:

$$F = n(n-1)/\sum_i n_i(n_i-1) \pm s$$

wherein:
F is the estimated size of the library
n is the total number of sequences obtained by sequencing,
ni is the number of sequence in the ith contig,
s is the standard error.

26 A method as claimed in claim 25 wherein an optimised library is generated by the further steps of:
(ix) providing a restriction site frequency for enzymes not used in step (i), optionally using the sequence information obtained at step (vii),
(x) selecting further restriction enzymes on the basis of restriction site frequency to generate a desired size of partition using the formula given in claim 23,
(xi) producing an optimised nucleic library in accordance with steps (i)-(vi) using at least one of these further restriction enzymes,
(xii) optionally repeating steps (vii) to (xi) until the desired library size is obtained.

27 A method as claimed in any one of claims 1 to 16 wherein the size of the sample nucleic acid is unknown,
and the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) are selected to produce the desired library size in

accordance with an assumed restriction site frequency of each enzyme in the sample nucleic acid.

28 A method as claimed in claim 27 wherein the restriction site frequency within the sample is assumed based on sequence information from the sample.

29 A method as claimed in claim 28 wherein the restriction site frequency is assumed to be randomly distributed

30 A method as claimed in any one of claims 27 and 29 wherein three 4nt- and one 6nt- cutting restriction enzymes are used in step (i).

31 A method as claimed in claim 30 wherein HpaII, AluI, DraI, and PstI are used in step (i).

32 A method as claimed in any one of claims 27 to 31 wherein the size of the resulting library is estimated by the further steps of:
(vii) sequencing the fragments in a fraction of the host cells in said library,
(viii) estimating the size of the library using formula:

$$F = n(n-1)/\sum_i n_i(n_i-1) \pm s$$

wherein:
F is the estimated size of the library
n is the total number of sequences obtained by sequencing,
ni is the number of sequence in the ith contig,
s is the standard error.

33 A method as claimed in claim 32 wherein the size of the sample is estimated by the further steps of:
(ix) providing the restriction site frequency of the enzymes used in step (i), optionally using the sequence information obtained at step(vii),
(x) calculating the sample size G using the formula:

$$N_{x1\sim x2} = GP_1^2 \sum_{k=x1}^{k=x2} \prod_{i=1}^{i} (1-P_i)^k$$

wherein:
Nx1~x2 is the number of fragments with length between x1 and x2
k is fragment length
x1 and x2 are upper and lower limits of the size range of the fragments in the library
Pi is the probability of having a restriction site at any given base for the 'i'th enzyme,

34 A method as claimed in claim 33 wherein an optimised library is generated by the further steps of:
(xi) providing a restriction site frequency for enzymes not used in step (i), optionally using the sequence information obtained at step(vii),
(xii) selecting further restriction enzymes on the basis of restriction site frequency to generate a desired size of partition using the formula given in claim 33,
(xiii) producing an optimised nucleic library in accordance with steps (i)-(vi) using at least one of these further restriction enzymes,
(xiv) optionally repeating steps (vii) to (xiii) until the desired library size is obtained.

35 A method as claimed in any one of the preceding claims wherein the sample nucleic acid comprises nucleic acid from two or more different sources which are pooled to produce a library comprising fragments from each.

36 A method for identifying a limited population of markers in a sample nucleic acid,
which method comprises:
(a) providing sample nucleic acid from at least two different sources,
(b) providing a library containing a representative partition of the sample nucleic acid in accordance with any one of claims 1 to 35,
(c) identifying differences within corresponding sequences from said different sources contained within the library

37 A method as claimed in claim 36 wherein the two different

nucleic sources are taken from different individuals.

38 A method as claimed in claim 36 wherein the markers are Single Nucleotide Polymorphisms.

39 A method as claimed in any one of claims 1 to 38 wherein the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) are selected in accordance with the output of program code run on a digital computer,

which computer comprises a processor, a data storage system, at least one input device, and at least one output device,

and which program code operates on the input of one or both of:

(i) a reference sequence or restriction map from the sample nucleic acid,

(ii) a preference regarding partition size, and optionally preferred region of the sample to include in the partition.

40 A method as claimed in claim 39 wherein the program code includes a look up table including reference restriction site target sequences for different 4 and 6nt cutting restriction enzymes.

41 A method as claimed in claim 39 wherein the program code performs a function in accordance with a formula described in claim 32 or claim 33.

42 A system for selecting the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) of the method of any one of claims 1 to 38,

which system comprises program code run on a digital computer,

which computer comprises a processor, a data storage system, at least one input device, and at least one output device,

and which program code operates on the input of one or both of:

(i) a reference sequence or restriction map from the sample nucleic

acid,
(ii) a preference regarding partition size, and optionally preferred region of the sample to include in the partition.

43 A system as claimed in claim 42 wherein the program code includes a look up table including reference restriction site target sequences for different 4 and 6nt cutting restriction enzymes.

44 A system as claimed in claim 43 wherein the program code performs a function in accordance with a formula described in claim 32 or claim 33.

45 A computer program for selecting the number of and type of the different restriction enzymes used in step (i), and the sub-set of layers selected in step (iii) of the method of any one of claims 1 to 41,

which computer program code operates on the input of one or both of:
(i) a reference sequence or restriction map from the sample nucleic acid,
(ii) a preference regarding partition size, and optionally preferred region of the sample to include in the partition,

and wherein the program code includes a look up table including reference restriction site target sequences for different 4 and 6nt cutting restriction enzymes,

and wherein the program code performs a function in accordance with a formula described in claim 32 or claim 33.

46 A computer program as claimed in claim 45 which is stored on a storage media or device readable by a general or special purpose programmable computer.

47 A process for producing a chip for use in assaying a limited population of polymorphisms within a sample, which process comprises:
(i) providing a population of probe sequences, which probe sequences are derived from a representative partition of sample nucleic acid provided in accordance with any one of claims 1 to 39,

and contain the population of polymorphisms,
(ii) incorporating the probe sequences into the chip.

48 A chip obtainable by the method of claim 47.

49 A method of genotyping a nucleic acid sample from an individual, which method comprises:
(i) providing the chip of claim 47 or claim 48,
(ii) isolating a representative partition of sample nucleic acid from the individual in accordance with the method used to provide the representative partition containing the population of polymorphisms contained in the probe sequences,
(iii) contacting the chip with the sample and determining hybridization of the sample nucleic acid thereto.